NOTES FOR SUMMER STATISTICS INSTITUTE COURSE

COMMON MISTAKES IN STATISTICS – SPOTTING THEM AND AVOIDING THEM

Day 2: Important Details about Statistical Inference

"The devil is in the details" (anonymous)

MAY 22-25, 2017

Instructor: Mary Parker

CONTENTS OF DAY 2

I. Random Sampling

In practice in applying statistical techniques, we're interested in *random variables* defined on the population under study.

Recall the examples mentioned yesterday:

1. In a medical study, the population might be all adults over age 50 who have high blood pressure.
2. In another study, the population might be all hospitals in the U.S. that perform heart bypass surgery.
3. If we're studying whether a certain die is fair or weighted, the population is all possible tosses of the die.

In these examples, we might be interested in the following random variables:

Example 1: The difference in blood pressure with and without taking a certain drug.

Example 2: The number of heart bypass surgeries performed in a particular year, or the number of such surgeries that are successful, or the number in which the patient has complications of surgery, etc.

Example 3: The number that comes up on the die.

Distinctions and Notation:

If we take a *sample of units from the population*, we have a corresponding *sample of values of the random variable*.

In Example 1:

- The *random variable* is "difference in blood pressure with and without taking the drug."
  - Call this random variable Y (upper case Y)
- The sample of *units from the population* is a sample of adults over age 50 who have high blood pressure.
  - Call them person 1, person 2, etc.
- The corresponding sample of values of the random variable will consist of values we will call $y_1, y_2, ..., y_n$ (lower case y's), where
  - n = number of people in the sample;
  - y_1 = the difference in blood pressures (that is, the value of Y) for the first person in the sample;
  - y_2 = the difference in blood pressures (that is, the value of Y) for the second person in the sample;
  - etc.

We can look at this another way, in terms of n random variables $Y_1, Y_2, ..., Y_n$, described as follows:

- The random process for Y_1 is "pick the first person in the sample"; the value of Y_1 is the value of Y for that person – i.e., y_1.
- The random process for Y_2 is "pick the second person in the sample"; the value of Y_2 is the value of Y for that person – i.e., y_2.
- etc.

The difference between using the small y's and the large Y's is that *when we use the small y's, those are actual values we found, where, with the large Y's we are still thinking of them as random variables which will change when we choose a different sample from the same population by the same probability-based (random) process.*

Note: Because the Y_i's are random variables, each has a distribution. Whether they are considered to have exactly the same distribution or "almost the same" distribution, is the topic of the next few slides.

Different types of random sampling

Some of the main types of random (probability-based) sampling that we can use are the following.
(*While not a comprehensive list, this is enough to illustrate the differences and their effect on the analyses.)*

1. Simple random sample from a finite population without replacement.
   (*What we usually are thinking about*.)

2. Simple random sample from a finite population with replacement.
   (*How strange! Why would anyone want to do that*?)

3. Simple random sample from an infinite population
   (*Tossing a die six times, etc*.)

4. Stratified random sample from a finite population
   (*Divide the population into strata, such as in Example 2, teaching hospitals and non-teaching hospitals. Take a simple random sample from each stratum, and combine those into one sample from the population.)*

Now, back to the mathematics . . .

The easiest processes to analyze mathematically are those in which the individual random variables $Y_1, Y_2, ... , Y_n$ are independent.

Intuitively speaking, "*independent*" means that *the values of any subset of the random variables $Y_1, Y_2, ... , Y_n$ do not influence the probabilities of the values of the other random variables in the list*.

Note: Because the Y_i's are random variables, each has a distribution. It would also be convenient, mathematically, if those random variables had the same probability distribution, which we call *identically distributed* (in this example, the distribution of Y).

So, in which of our four types of random samples are the Y_i's independent and identically distributed (denoted iid) ?

1. Simple random sampling without replacement. **Not here.** Suppose the population has 1000 members in it. When we choose the value for Y_1 then there are 999 members left in the population from which we will choose Y_2.

   So we are not choosing Y_2 from exactly the same population as we chose Y_1. We can also see that Y_2 is not independent of Y_1 because knowing the value for Y_1 gives us some information about the value for Y_2 .

2. Simple random sampling with replacement. YES, these are iid.

   Suppose the population has 1000 members in it. We choose the value for Y_1, record that value as y_1, and then replace that value in the population before we draw a value for Y_2 .

   So we are drawing Y_2 from the same population we drew Y_1 from, and the value we chose for Y_1 has no impact on the value we chose for Y_2 .

Let's pause here. *Now we see why someone might choose to do this strange method of sampling with replacement. It makes the mathematics easier!*

3. Simple random sample from an infinite population

   What, exactly does an infinite population mean?
   That means that our random variables Y_i are values from a particular probability distribution. So when we take a value for Y_1 from that population, we next take the value Y_2 from the same population. So here the Y_i values are independent and identically distributed.
   (Tossing a fair die: the population has a discrete uniform distribution on the set of whole numbers from 1 to 6.)

4. Stratified random sampling.

   Considering this gives us no new insights into this situation about independent and identically distributed random variables Y_i , so we will postpone discussion of it for now.

Which of our applied examples are iid?

Recall Example 3 above: We toss a die; the number that comes up on the die is the value of our random variable Y.

- In terms of the preliminary definition:
    - The population is all possible tosses of the die.
    - A simple random sample is n different tosses.
- The different tosses of the die are *independent events* (i.e., what happens in some tosses has no influence on the other tosses), which means that in the precise definition above, the random variables $Y_1, Y_2, \ldots, Y_n$ are indeed independent: The numbers that come up in some tosses in no way influence the numbers that come up in other tosses.

Recall Examples 1 and 2 above:

Example 1: Choose a random sample of adults over 50 with high blood pressure.

Example 2: Choose a sample of hospitals in the US that perform heart bypass surgery.

No one explicitly said whether the sampling would be with replacement or without replacement, but we suspect that those doing the sampling would do it without replacement. But we just learned that the mathematics is a lot easier if we sample with replacement. What happens? Is this a major problem?

Mathematics can come to the rescue!

What mathematics?

The corrections needed to the usual formulas are multipliers to lower the variance appropriately.

For single means and single proportions, they are something like $\sqrt{1-\frac{n}{N}}$ where n is the sample size and N is the population size.

Try for yourself to see how large a fraction of the population you need to sample to get enough reduction in the variance when estimating a single mean or single proportion to be worth correcting.

More explanation of this is available at https://www.ma.utexas.edu/users/parker/sampling/repl.htm

How to avoid mistakes when sampling without replacement and using the standard formulas in applied statistics courses.

Pay attention to what percentage of the population you are sampling. If it is a small percentage, don't worry about it.

If it is a larger percentage (more than 5% or 10%) then congratulate yourself on getting quite a lot of information about your population! And then find the finite correction factors needed to adjust your variances to "get credit in your calculations" for the extra information you have!

Textbook definitions of simple random sampling

Generally speaking, when textbooks discuss types of sampling, their definition of a simple random sample is essentially this:

> A simple random sample of size n is a sample taken from the population in such a way that every possible subset of the population of size n has an equal probability of being chosen.

Notice that definition assumes sampling with replacement, so without the iid assumption.

Generally speaking, when those same textbooks discuss formulas, and certainly when they discuss deriving formulas, they start with

> Consider a random sample of size *n* from a probability distribution.

Notice that this assumes sampling from an infinite population, which means we have the iid assumption.

No wonder people finishing these courses are not completely clear on what to do!

What should we be doing to clarify the situation?

In elementary statistics courses?

____

In this course?

____

"Simple random sample" in the rest of this course

For the rest of this course, when I say that our data came from a simple random sample, here's what I mean.

- If it is relating to an applied problem (where the sampling is from a finite population) I assume the definition of simple random sample given earlier – which includes sampling without replacement.
- Even though it isn't quite correct, I will use the usual formulas that come from assuming independent samples.
- I will assume that one of the conditions of our model (without stating it) is that the sample is no larger than 10% of the population.

When I do this, I am uncomfortable as a mathematician. But, as a statistics teacher, I recognize that this is the way almost all textbooks are written, so it seems appropriate here as well.

Beyond Simple Random Sampling

Stratified Random Sampling

- This is used to get a “more reliably representative” sample than simple random sampling.
- If the stratification is done in a useful manner, you have a smaller variance in the sample statistic.

Why would “more reliably representative” translate to a smaller variance in the sample statistic? What is the variance in the sample statistic measuring?

How much smaller?

It depends on how useful your stratification is. The formulas are more complex than those for simple random sampling, but not unreasonably complex.

II. WHY RANDOM SAMPLING IS IMPORTANT

Recall the Myth:

"A random sample will be representative of the population".

A slightly better explanation (partly true but partly Urban Legend):

"Random sampling prevents bias by giving all individuals an equal chance to be chosen."

- The element of truth: Random sampling *does* eliminate *systematic* bias.
- A practical rationale: This explanation is often the best plausible explanation that is acceptable to someone with little mathematical background.
- However, this statement could be misinterpreted as the myth above.

An additional, very important, reason why random sampling is important, at least in frequentist statistical procedures, which are those most often taught (especially in introductory classes) and used:

The Real Reason: *Only a probability-based sampling method allows us to compute the probabilities needed for statistical inference (confidence intervals and hypothesis tests.)*

The next section elaborates.

III. OVERVIEW OF FREQUENTIST HYPOTHESIS TESTING

Type of Situation where a Hypothesis Test is used:

- We *suspect* a certain pattern in a certain situation.
- But we realize that natural variability or imperfect measurement might produce an apparent pattern that isn't really there.

Basic Elements of Most Frequentist Hypothesis Tests:
Most commonly-used ("parametric"), frequentist hypothesis tests involve the following four elements:

i. *Model assumptions*

ii. *Null and alternative hypotheses*

iii. A *test statistic*

This is something that
  a. Is *calculated by a rule from a sample*;
  b. Is a measure of the strength of the pattern we are studying; and
  c. Has the property that, *if* the null hypothesis is true, *extreme values of the test statistic are rare, and hence cast doubt on the null hypothesis*.

iv. A *mathematical theorem* saying,

"<u>If</u> the model assumptions <u>and</u> the null hypothesis are <u>both</u> true, then the *sampling distribution* of the test statistic has a certain particular form."

Note:

- The *sampling distribution* is the probability distribution of the test statistic, when considering *all* possible suitably random samples of the same size. (More later.)
- The exact details of these four elements will depend on the particular hypothesis test.
- In particular, the form of the sampling distribution will depend on the hypothesis test.

Illustration: Large Sample z-Test for the mean, with two-sided alternative

The above elements for this test are:

1. *Model assumptions*: We are working with <u>simple random samples</u> of a random variable Y that has a <u>normal distribution</u> with known standard deviation.

2. *Null hypothesis*: “The mean of the random variable Y is a certain value μ_0.”

Alternative hypothesis: "The mean of the random variable Y is not μ_0." (This is called the *two-sided alternative*.)

3. *Test statistic*: $\bar{y}$ (the *sample mean* of a simple random sample of size n from the random variable Y).

Before discussing item 4 (the mathematical theorem), we first need to:

1. Clarify terminology
2. Discuss sampling distributions

1. **Terminology and common confusions:**

- The ***mean of the random variable*** Y is also called the *expected value* or the *expectation* of Y.
  - It’s denoted E(Y).
  - It’s also called the *population mean*, often denoted as μ.
  - It’s what we do *not* know in this example.

- A ***sample mean*** is typically denoted $\bar{y}$ (read "y-bar").
  - It’s calculated from a sample $y_1, y_2, \ldots, y_n$ of values of Y by the familiar formula $\bar{y} = (y_1 + y_2 + \ldots + y_n)/n$.

- The sample mean $\bar{y}$ is an *estimate* of the population mean μ, but they are usually *not* the same.
  - Confusing them is a **common mistake**.

- Note the articles: "*the* population mean" but "*a* sample mean".
  - There is only one *population* mean associated with the random variable Y.
  - However, a *sample* mean depends on the sample chosen.
  - Since there are many possible samples, there are many possible sample means.

Illustration:

https://istats.shinyapps.io/sampdist_cont/

(Use “Bell-shaped” and explore.)

Putting this in a more general framework:

- A *parameter* is a constant associated with a population.
  - In our example: The population mean μ is a parameter.
  - When applying statistics, parameters are usually unknown.
  - However, the goal is often to gain some information about parameters.
- To help gain information about (unknown) parameters, we use *estimates* that are calculated from a sample.
  - In our example: We calculate the sample mean $\bar{y}$ as an estimate of the population mean (parameter) μ.

"Variance" is another common example where parameter and estimate might be confused:

- The *population variance* (or "the variance of Y") is a parameter, usually called σ^2 (or Var(Y)).
- If we have a sample from Y, we can calculate the *sample variance*, usually called s^2.
- A sample variance is an estimate of the population variance.
- Different samples may give different estimates.
- *Confusing population and sample variance is a* ***common mistake.***

Mental shift in measuring variability

- We will soon be seriously looking at pictures of distributions and discussing the center and variability of the distributions.

- Variances are a common measure of variability when doing theoretical calculations to derive the various important results, because, in various important situations, the variance of a sum of random variables is the sum of the variances of the individual random variables.

- Consider, however, that if the variable is the heights of people in inches, then the mean is measured in inches, but the variance is measured in square inches.

- The discrepancy between the units of measurement of the mean and variance makes it hard to think about all of this in the context of the situation.

- Also , when we start to look at pictures of distributions, it is more convenient to have the mean and the measure of variance have the same units.

- Thus, the standard deviation is defined to be the square root of the variance.   So the standard deviation of the heights in our example is also measured in inches.

2. **Sampling Distribution:**

Although we apply a hypothesis test using a single sample, we need to step back and consider *all* possible suitably random samples of Y of size n, in order to understand the test. In our example:

- For each simple random sample of Y of size n, we get a value of $\bar{y}$.
- We thus have a *new* random variable $\bar{Y}_n$:
  - The associated random process is "pick a simple random sample of size n"
  - The *value* of $\bar{Y}_n$ is the *sample mean* $\bar{y}$ for this sample
- Note that
  - $\bar{Y}_n$ stands for the new random variable
  - $\bar{y}$ stands for the value of $\bar{Y}_n$, for a particular sample of size n.
  - $\bar{y}$ (the value of $\bar{Y}_n$) depends on the sample, and typically varies from sample to sample.
- The distribution of the new random variable $\bar{Y}_n$ is called the *sampling distribution of* $\bar{Y}_n$ (or the *sampling distribution of the mean*).
- Note: $\bar{Y}_n$ is an example of an *estimator*: a random variable whose values are estimates.

Now we can state the theorem that the large sample z-test for the mean relies on:

4. The *theorem* states: *If* the model assumptions are all true (i.e., *if* Y is normal *and* all samples considered are simple random samples), and *if in addition* the mean of Y is indeed μ_0 (i.e., *if* the null hypothesis is true), then

   i. The sampling distribution of $\bar{Y}_n$ is normal

   ii. The sampling distribution of $\bar{Y}_n$ has mean μ_0

   iii. The sampling distribution of $\bar{Y}_n$ has standard deviation $\sigma/\sqrt{n}$, where σ is the standard deviation of the original random variable Y.

Check that this is consistent with what the simulation shows.

https://istats.shinyapps.io/sampdist_cont/
Choose "Bell Shaped" and try sample size n = 25.

Also note:

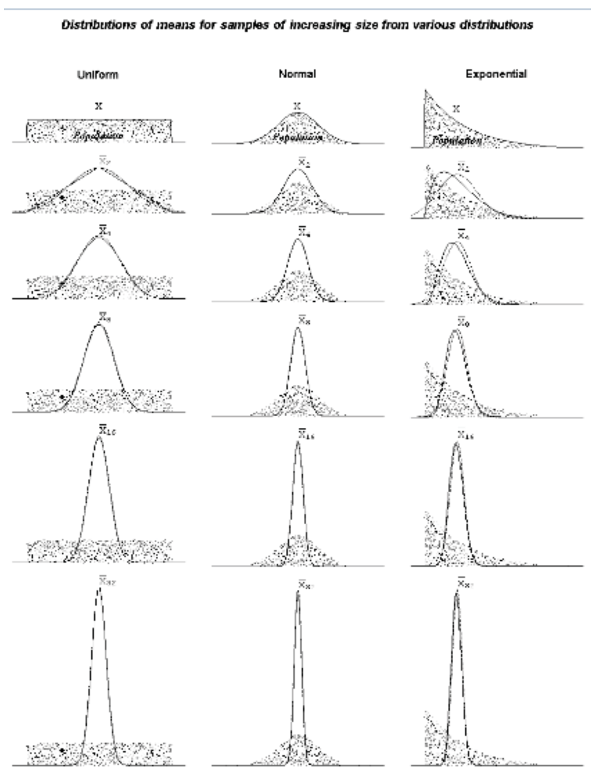
- $\sigma/\sqrt{n}$ is smaller than σ (if n is larger than 1)
- The larger n is, the smaller $\sigma/\sqrt{n}$ is.
- ***Why is this nice?***

More Terminology: σ is called the *population standard deviation* of Y; it is *not* the same as the *sample standard deviation* s, although s is an estimate of σ.

The following pictures and chart summarize the conclusion of the theorem and related information about robustness.

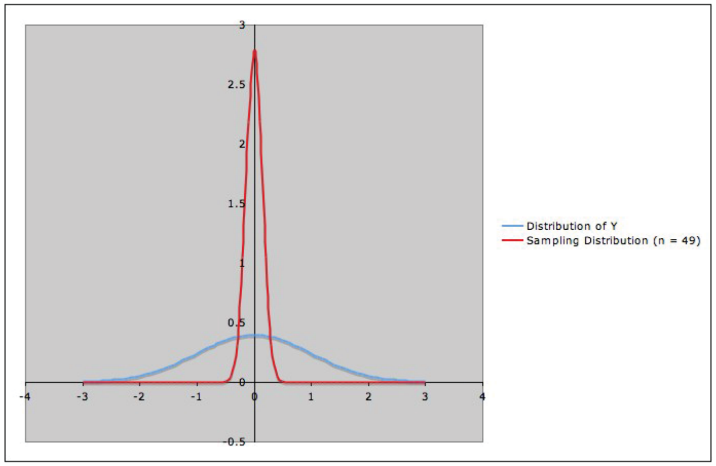
http://www.austincc.edu/mparker/1342/cltdemos.htm and look at the middle column.

How robust is the normality of the sample mean against deviations from the population dist'n being normal?   Answer: Very. See other pictures here.



	Considering the population ↓	Considering *one* sample ↓	Considering *all* suitable samples ↓
	Random variable Y (population distribution)	Related quantity calculated from a sample $y_1, y_2, \ldots, y_n$	**Random variable $\overline{Y}_n$ (sampling distribution)**
Type of Distribution	Y *has* a normal distribution	The sample is *from* the (normal) distribution of Y	$\overline{Y}_n$ *has* a normal distribution
Mean	*Population mean* μ ($\mu = \mu_0$ if null hypothesis true)	*Sample mean* $\bar{y}$ = $(y_1 + y_2 + \ldots y_n)/n$ $\bar{y}$ is an estimate of the population mean μ	*Mean of the sampling distribution of* $\bar{y}$ -- it's also μ. ($\mu = \mu_0$ if null hypothesis true)
Standard deviation	*Population standard deviation* σ	*Sample standard deviation* s = $\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(\bar{y}-y_i)^2}$ s is an estimate of the population standard deviation σ	*Sampling distribution standard deviation* $\sigma/\sqrt{n}$

↑ From the Theorem



- Which column of the chart on the previous page corresponds to the blue distribution?
- Which column of the chart on the previous page corresponds to the red distribution?
- How could you tell without looking at the legend?

The roles of the model assumptions for this hypothesis test (large sample z-test for the mean)**:**

Recall:

The theorem has *three assumptions*:

Assumption 1: Y has a normal distribution (*a model assumption)*.

Assumption 2: All samples considered are simple random samples (*also a model assumption)*.

Assumption 3: The null hypothesis is true (*assumption for the theorem, but not a model assumption)*.

The theorem also has *three conclusions*:

Conclusion 1: The sampling distribution of $\overline{Y}_n$ is normal

Conclusion 2: The sampling distribution of $\overline{Y}_n$ has mean μ_0

Conclusion 3: The sampling distribution of $\overline{Y}_n$ has standard deviation $\sigma/\sqrt{n}$, where σ is the standard deviation of the original random variable Y.

The following chart shows *which assumptions each conclusion depends upon*:

	Assumptions of Theorem ↓		
Conclusions ***about Sampling Distribution*** (Distribution of $\bar{Y}_n$) ↓	1. Y normal (model assumption)	2: Simple random samples (model assumption)	3: Null hypothesis true (not a model assumption)
1: Normal	√	√	
2: Mean μ_0			√
3: Standard deviation $\sigma/\sqrt{n}$		√	

↑
Corresponds to third column in table on the previous page

Note that the model assumption that the sample is a simple random sample (in particular, that the Y_i's as defined earlier are independent) *is used to prove:*

1. that the sampling distribution is normal *and*
2. (even more importantly) that the standard deviation of the sampling distribution is $\sigma/\sqrt{n}$.

<u>This illustrates a general phenomenon that *independence conditions are usually important in deriving the variance formulas usually taught in applied statistics courses.*</u>

Consequences of model assumptions:

1. *<u>If</u>* the conclusion of the theorem is true, the sampling distribution of $\bar{Y}_n$ is narrower than the original distribution of Y

- In fact, conclusion 3 of the theorem gives us an idea of just how narrow it is, depending on n.
- *This will allow us to construct a useful hypothesis test.*

2. The only way we know the conclusion is true is if we know the hypotheses of the theorem (the model assumptions and the null hypothesis) are true.

3. Thus: *If the model assumptions are <u>not</u> true, then we <u>do not know</u> that the theorem is true, so we <u>do not know</u> that the hypothesis test is valid.*

In the example (large sample z-test for a mean), this translates to:

If the sample is not a simple random sample, <u>or</u> if the random variable is not normal, then the mathematical reasoning establishing the validity of the test breaks down.

HOWEVER, it is often true that a test can be either used as is, or slightly modified, to allow for some of the assumptions not being exactly met. The discussion of this is described as the "robustness" of the test.

Criteria for robustness are developed in various ways.
1. Different (often related) mathematical reasoning than that used to develop the original test.
2. Simulation studies.

Robustness of the Large-Sample z-test (Central Limit Thm, etc.)

What if we don't know that the population is normally distributed with known standard deviation?

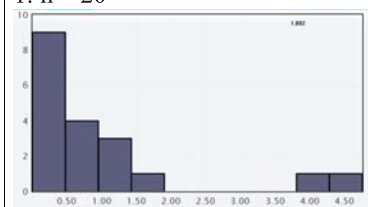
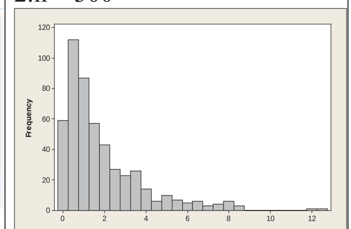
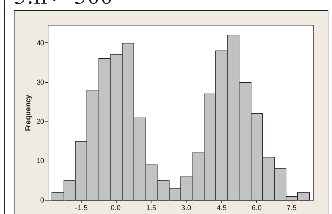
1) Do not know that the pop'n is normally distributed.
   Solution: The **Central Limit Theorem** tells us that, if the sample size is large enough, the distribution of the sample mean is approximately normal. Large-sample z-test applies when $n > 30$. (Sample size needed depends on shape of pop'n dist'n.)

2) Do know that the population is normally distributed but don't know the standard deviation of the population.
   Solution: The t-distribution procedures were developed to address this situation.

3) Don't know that the population is normally distributed and don't know the standard deviation of the population.
   Solution: The t-distribution procedures are robust against deviations from normality. In practice, that means something like this:
   a) Large sample size: In many situations, the t-distribution procedures can be used. If the sample data shows substantial skewness then larger samples are needed to be OK.
   b) Medium sample size ($15<n<40$) If the sample data doesn't show strong skewness, then the t-procedures can fairly safely be used.
   c) Small samples ($n < 16$) Use t-procedures only if you have some reason to believe the population is quite close to normally distributed.

Important lesson here: When the model assumptions aren't exactly met by your data, think carefully about how they aren't met and whether the procedure is robust against deviations from those assumptions.

Quiz:
QUIZ: Graphs given are histograms of the sample. Would this test be reasonable to use in the following situations? Why or why not?

1. $n = 20$ 	2. $n > 300$ 
3. $n = 1017$ The sample consists of people who respond to a survey on a website.	3. $n > 300$ 
5. $n = 8$ Y is a random variable that can only take on values equal to 0 or 1.	6. $n = 100$. Y is a random variable that can only take on values equal to 0 or 1.

Importance of Model Assumptions in General:

Different hypothesis tests have different model assumptions.

- Some tests apply to random samples that are not simple.
- For many tests, the model assumptions consist of *several* assumptions.
- If *any one* of these model assumptions is not true, we *do not know* that the test is valid. (*The bad news)*

Many techniques are ***robust*** *to some departures from at least some model assumptions*. (*The good news)*

- "Robustness" means that if the particular assumption is *not too far* from true, then the technique is still approximately valid.
- As we saw, the large sample z-test for the mean is robust to departures from normality. In particular, for *large enough* sample sizes, the test is very close to accurate.
    - o How large is large enough depends on the distribution of the random variable Y.
    - o This can be explored by simulation and re-sampling techniques.
- Robustness depends on the particular procedure; there are no "one size fits all" rules.

Caution re terminology: "Robust" is used in other ways – for example, "a finding is robust" could be used to say that the finding appears to be true in a wide variety of situations, or that it has been established in several ways.

A **common mistake** in using statistics*: Using a hypothesis test without paying attention to whether or not the model assumptions are true and whether or not the technique is robust to possible departures from model assumptions is a.*

- ***The process of peer review often does not catch these and other statistical mistakes***.
- This is one of many reasons why the "results" published in peer-reviewed journals are often false or not well substantiated.
- Compounding the problem, journals rarely retract articles with mistakes. For examples and discussion, see
    - o Allison et al, A Tragedy of Errors, Nature 530, 4 February 2016, 28 – 29, http://www.nature.com/news/reproducibility-a-tragedy-of-errors-1.19264
    - o Kamoun, S. and C. Zipfel, Scientific record: Class uncorrected errors as misconduct, Nature 531, 10 March 2016, 173, http://www.nature.com/nature/journal/v531/n7593/full/531173e.html
    - o Gelman and commenters, http://andrewgelman.com/2016/02/22/its-too-hard-to-publish-criticisms-and-obtain-data-for-replication/
- Resources for finding (at least some) retractions and other reports of errors include
    - o PubPeer (https://pubpeer.com/)
    - o Retraction Watch (http://retractionwatch.com/)
    - o Authors sometimes post errors or corrections on their own website.
    - o bioRxiv (http://biorxiv.org/) is a biology preprint server that also allows moderated public comments on posted papers. It classifies replication studies as Confirmatory or Contradictory.

IV: FREQUENTIST CONFIDENCE INTERVALS

Before continuing the discussion of hypothesis tests, it will be helpful first to discuss the related concept of confidence intervals.

The General Situation:

- We're considering a *random variable* Y.
- We're interested in a certain *parameter* (e.g., a proportion, or mean, or regression coefficient, or variance) associated with the random variable Y (i.e., associated with the population)
- We *don't* know the value of the parameter.
- *Goal 1*: We'd like to estimate the unknown parameter, using data from a sample. (A ***point estimate***)
- *But since estimates are always uncertain, we also need:*
- *Goal 2*: We'd like to get some sense of how good our estimate is. (Typically achieved by an ***interval estimate***)

The first goal is usually easier than the second.

Example: If the parameter we're interested in estimating is the *mean of the random variable* (i.e., the population mean, which we call ____), we can estimate it using a *sample mean* (which we call ____).

The rough idea for achieving the second goal (getting some sense of how good our estimate is):

> We'd like to get a *range of plausible values for the unknown parameter*. ("What we want.") We will call this range of plausible values a *<u>confidence interval</u>.*

The usual method for calculating a confidence interval has a lot in common with hypothesis testing:

- It involves the sampling distribution
- It depends on model assumptions.

A little more specifically:

- Although we typically have just one sample at hand when we do statistics, *the reasoning used in classical frequentist inference depends on thinking about <u>all possible</u> suitable samples of the same size n.*
- Which samples are considered "suitable" will depend on the particular statistical procedure to be used.
- Each confidence interval procedure has *model assumptions* that are needed to ensure that the reasoning behind the procedure is sound.
- The model assumptions determine (among other things) which samples are "suitable."
- The procedure is applicable only to "suitable" samples.

Illustration: Large Sample z-Procedure for a Confidence Interval for a Mean

- The parameter we want to estimate is the population mean $\mu = E(Y)$ of the random variable Y.
- The model assumptions for this procedure are: The random variable is <u>normal</u>, and samples are <u>simple random samples</u>.
    - So in this case, "suitable sample" means "simple random sample".
    - For this procedure, we also need to know that Y is normal, so that both model assumptions are satisfied.

Notation and terminology:

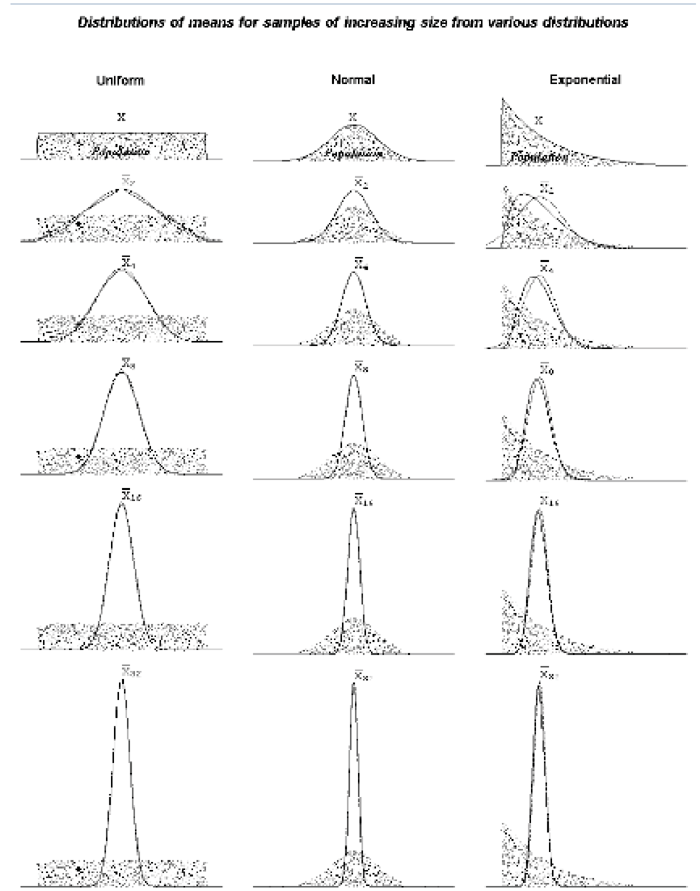
- We'll use σ to denote the (population) standard deviation of Y.
- We have a simple random sample, say of size n, consisting of observations $y_1, y_2, \ldots, y_n$.
    - For example, if Y is "height of an adult American male," we take a simple random sample of n adult American males; $y_1, y_2, \ldots, y_n$ are their heights.
- We use the sample mean $\bar{y} = (y_1 + y_2 + \ldots + y_n)/n$ as our estimate of the population mean μ.
    - This is an example of a ***point estimate*** -- a numerical estimate with no indication of how good the estimate is.

More details:

- To get an idea of how good our estimate is, we use the concept of *confidence interval.*
    - This is an example of an ***interval estimate***.
- To understand the concept of "confidence interval" in detail, we need to consider *<u>all possible</u> simple random samples of size n from Y*.
    - In the specific example, we consider all possible simple random samples of n adult American males.
    - For each such sample, the heights of the men in the sample of people constitute our simple random sample of size n from Y.
- We consider the *sample means* $\bar{y}$ *for <u>all possible</u> simple random samples of size n from Y.*
    - This amounts to defining a new random variable, which we will call $\bar{Y}_n$ (read "Y-bar sub n", or "Y-sub-n- bar").
    - We can describe the random variable $\bar{Y}_n$ briefly as "sample mean of a simple random sample of size n from Y", or more explicitly as: "pick a simple random sample of size n from Y <u>and</u> calculate its sample mean".
    - Note *that each value of* $\bar{Y}_n$ is an estimate of the population mean μ.
        - e.g. each simple random sample of n adult American males gives us an estimate of the population mean μ.

- This new random variable $\overline{Y}_n$ has a distribution, called a *sampling distribution* (since it arises from considering varying samples).
    - The values of $\overline{Y}_n$ are all the possible values of sample means $\overline{y}$ of simple random samples of size n of Y – i.e, the values of our *estimates* of μ.
    - *The sampling distribution (distribution of* $\overline{Y}_n$*) gives us information about the* <u>*variability*</u> *(as samples vary) of our estimates* of the population mean μ.
    - See that illustrated on our page of pictures of sampling distributions of sample means.

http://www.austincc.edu/mparker/1342/cltdemos.htm and look at the middle column illustrating when Y has a normal distribution.



- A mathematical theorem tells us that *if* the model assumptions are true, then:

  1. The sampling distribution is normal
  2. The mean of the sampling distribution is also μ.
  3. The sampling distribution has standard deviation $\sigma / \sqrt{n}$

- Use these conclusions to compare and contrast the shapes of the distribution of Y and the distribution of $\bar{Y}_n$

  - What is the same? ________________________
  - What is different? ________________________
  - How do the standard deviations compare? ___________________________________

- The chart on the next page (best read one column at a time) and picture below summarize some of this information.